DOCUMENT RESUME

ED 382 639 TM 023 069

AUTHOR McClain, Andrew L.
TITLE Effect Size as an Alternative to Statistical Significance Testing.
PUB DATE Apr 95
NOTE 18p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, April 18-22, 1995).
PUB TYPE Information Analyses (070) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Correlation; *Effect Size; Research Methodology; *Sample Size; *Statistical Significance; *Testing
IDENTIFIERS Scattergrams; *Variance (Statistical)

ABSTRACT

The present paper discusses criticisms of statistical significance testing from both historical and contemporary perspectives. Statistical significance testing is greatly influenced by sample size and often results in meaningless information being over-reported. Variance-accounted-for-effect sizes are presented as an alternative to statistical significance testing. A review of the "Journal of Clinical Psychology" (1993) reveals a continued reliance on statistical significance testing on the part of researchers. Finally, scatterplots and correlation coefficients are presented to illustrate the lack of linear relationship between sample size and effect size. Two figures are included. (Contains 24 references.) (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 382 639

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

This document has been reproduced as received from the person or organization originating it.

Minor changes have been made to improve reproduction quality.

Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

ANDREW L. McCLAIN

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC).

Effect Size as an Alternative to Statistical Significance Testing

Andrew L. McClain

Texas A&M University 77845-4225

Paper presented at the annual meeting of the American Educational Research Association (session #52.26), San Francisco, CA, April 22, 1995.

BEST COPY AVAILABLE

Abstract

The present paper discusses common criticisms of statistical significance testing from both historical and contemporary perspectives. Statistical significance testing is greatly influenced by sample size and often results in meaningless information being over-reported. Variance-accounted-for effect sizes are presented as an alternative to statistical significance testing. A review of the Journal of Clinical Psychology (1993) reveals a continued reliance on statistical significance testing on the part of researchers. Finally, scatterplots and correlation coefficients are presented to illustrate the lack of linear relationship between sample size and effect size.

Effect Size as an Alternative to Statistical Significance Testing

Historically, the use of statistical significance testing for interpreting research results has generated considerable debate (Bakan, 1966; Carver, 1978, 1993; Huberty, 1987, 1993; Thompson, 1988, 1989a, 1989b, 1994). Articles on the limits of statistical significance testing have appeared in the American Psychologist (Cohen, 1990, 1994; Kupfersmid, 1988; Rosenthal, 1991; Rosnow & Rosenthal, 1989), the Journal of Consulting and Clinical Psychology (Meehl, 1978), and the Journal of Counseling Psychology (Atkinson, Furlong, & Wampold, 1982; Fagley & McKinney, 1983). In addition, the entire Summer, 1993, issue of the Journal of Experimental Education (Thompson, 1993) addresses issues related to statistical significance testing, effect sizes, and replicability in empirical research. The purposes of the present paper are to elaborate criticisms of the over-reliance on statistical significance testing and to present alternatives that may successfully augment the evaluation of statistical significance testing in psychological research.

It is indeed disturbing to take note of the published research that continues to rely so much on statistical significance testing. In his latest article on the limitations of statistical significance testing, Cohen (1994) writes,

> After 4 decades of severe criticism, the ritual of null hypothesis significance testing--mechanical dichotomous decisions around a sacred .05 criterion--still persists.

> . . . I argue herein that null hypothesis significance testing has not only failed to support the advance of psychology as a science but also has seriously impeded it. (p. 997)

Similarly, Bakan (1966) states, ". . . the test of significance does not provide the information concerning psychological phenomena characteristically attributed to it; and that, furthermore, a great deal of mischief has been associated with its use" (p. 423). Statistical significance testing has been repeatedly criticized for its over-reliance on sample size and inability to detect meaningful results.

Researchers who have had the experience of working with large samples soon realize that virtually all null hypotheses will be rejected at some sample size, since ". . . the null hypothesis, taken literally, is always false" (Meehl, 1978, p. 822). Literally, statistical significance can be "achieved" at some given sample size; it simply becomes a matter of obtaining enough subjects. As Thompson (1992) remarks,

> Statistical significance testing can involve a tautological logic in which tired researchers, having collected data from hundreds of subjects, then conduct a statistical test to evaluate whether there were a lot of subjects, which the researchers already know, because they collected all the data and know they're tired. (p. 436)

In regard to the meaningfulness of results, the reliance on statistical significance testing has resulted in researchers

conducting *t*-tests and discarding *t*-values that do not meet the conventional level of statistical significance; in fact, the .05 level of significance has become somewhat of an arbitrary gatekeeper to scientific knowledge. Results not meeting this criteria are viewed as being trivial or unimportant (Greenwald, 1975). Rosnow and Rosenthal state (1989), ". . . surely God loves the .06 nearly as much as the .05. Can there be any doubt that God views the strength of evidence for or against the null as a fairly continuous function of the magnitude of *p*?" (p. 1277).

It seems that this dichotomous decision-making grew from Fisher's use of statistical significance testing in making decisions related to agriculture (e.g., whether to plant a crop of wheat or use manure to fertilize). In fact, many researchers cite Fisher as supporting an alpha of .05 as the criterion for statistical significance. It is interesting to note that Fisher did not set a fixed level of significance and called for researchers to evaluate individual cases in reference to data and theory (Huberty, 1993).

Clearly, this mechanistic, clear-cut (relatively mindless) method of making decisions offered a feeling of "scientific" objectivity that researchers found appealing (Cohen, 1990). Unfortunately, the cost for this "objectivity" is meaningful (although not statistically significant) information being under-reported and meaningless (statistically significant) information being over-reported.

One alternative to statistical significance testing is the use of variance-accounted-for effect sizes. These statistics are applicable because all conventional parametric analyses (e.g., *t*-tests, ANOVA, ANCOVA, regression) are correlational in nature or special cases of canonical correlation analysis (Knapp, 1978). Effect sizes allow the researcher to evaluate common variance shared between variables; in other words, how much of the variance in A can be explained by B? Unlike measures of statistical significance, "corrected" shared variance statistics (Snyder & Lawson, 1993) are not influenced by sample size thus allowing the researcher to form a clearer understanding of the relationships present in the data (Craig, Eison, & Metze, 1976; Thompson, 1989a).

Because effect sizes are correlational in nature, they capitalize on all variance available in the sample data, including sampling error. Sampling error by its very nature is variance that is unique to a given sample and does not occur at all in the population. An uncorrected effect size such as eta^2 may have a positive bias in representing shared variance due to capitalizing on sampling error. Corrected effect sizes such as $omega^2$ (Hays, 1973) and the Wherry formula (Wherry, 1931) adjust for the proportion of variance that is due to sampling error. Conceptually, a corrected effect size can equal but not exceed the magnitude of an uncorrected effect size (e.g., $omega^2 \leq eta^2$). For a review on the implications of using corrected and

uncorrected effect sizes, the reader is referred to Snyder and Lawson (1993).

The recently published edition of the Publication Manual of the American Psychological Association (1994) contains a section outlining the use of statistical significance testing in published research. Suggestions are made to report the exact probability (*p*) values of the computed test statistic; most statistical computer packages report these values. In addition, the term, "statistically", is used in conjunction with the term, "significant", to denote statistical significance testing; many researchers use only the term, "significant", in reporting statistically significant findings thus leading the reader to the erroneous conclusion that "significant" results are meaningful (Carver, 1993). The manual (APA, 1994) gives the following example, "The effect of age was not statistically significant, $F(1, 123) = 2.45$, $p = .12$" (p. 17).

In regard to reporting effect sizes in published research, the manual states:

> Neither of the two types of probability values reflects the importance or magnitude of an effect because both depend on sample size. You can estimate the magnitude of an effect with a number of measures that do not depend on sample size. . . . You are encouraged to report effect-size information. (APA, 1994, p. 18)

It seems that the new Publication Manual of the American Psychological Association may prove instrumental in encouraging

researchers to report effect sizes in published research. The present paper evaluates published research using the above mentioned criteria for the Journal of Clinical Psychology (1993). Corrected and uncorrected effect sizes are computed for both *t* and *F* statistics along a varying range of sample sizes and statistical significance levels.

Method

Sample

One-hundred-thirteen articles from the Journal of Clinical Psychology for the year 1993 were examined. Fifty-four (48%) of these articles contained either *F* and/or *t* values. One-hundred-twenty-eight test statistics were obtained. Thirty-three *t* values from independent *t*-tests and 48 *F* values from one-way ANOVA procedures were used to calculate effect sizes.

Procedure

Information related to reported test statistic, statistical analysis, sample size, reported effect size, *p*-calculated value, and degrees of freedom were obtained from 54 articles reporting either an *F* or *t* value. Where available, 2 *F* values and 2 *t* values were obtained from each article; care was taken to obtain one test statistic with the highest *p*-calculated value (e.g., $p = .50$) and the other test statistic with the lowest *p*-calculated value (e.g., $p = .0001$). In the optimal case, each article yielded 2 *F* values and 2 *t* values with varying *p*-calculated values. All test statistics (i.e., *F*'s and *t*'s) meeting these criteria were noted and examined.

For heuristic purposes, 33 *t* values from independent *t*-tests and 48 *F* values from one-way ANOVAs were used to compute both corrected (omega^2) and uncorrected (eta^2) effect sizes. Examples from independent *t*-tests and one-way ANOVA's were used to make application of effect size formulas more straightforward. In addition, some test statistics were not used due to ambiguous or missing information (e.g., sample size, degrees of freedom, etc.). Scatterplots were used to visually illustrate the relationship between corrected and uncorrected effect sizes and sample size.

Results

Of the 128 test statistics examined, 9 (7%) were reported with effect sizes. Effect sizes were reported in 5 out of 54 articles (9%). Of the 54 articles reviewed, 44 (81%) used the term "significant", 7 (13%) used both the terms "significant" and "statistically significant", and 3 (6%) used the term "statistically significant" in reporting statistical significance.

Figure 1 presents a scatterplot for corrected effect size (Snyder & Lawson, 1993) and sample size. Figure 2 presents a scatterplot for uncorrected effect size and sample size. The correlation between corrected effect size and sample size was -0.1007. The correlation between uncorrected effect size and sample size was -0.1189.

Discussion

In reviewing the articles, it appears that these researchers continue to rely heavily on statistical significance testing in evaluating research results. Unfortunately, only 5 researchers reported effect size information in their studies. In addition, the majority of articles (81%) referred to these statistically significant results as being "significant", thus implying that the results were meaningful. Carver (1993) emphasizes, "There is no good excuse for saying that a statistically significant result is significant because this language erroneously suggests to many readers that the result is automatically large, important, and substantial" (p. 288). One wonders if such language is responsible for perpetuating the myth of statistical significance testing.

The scatterplots presented in Figures 1 and 2 visually illustrate the lack of linear relationship between sample size and effect size for both corrected and uncorrected effect sizes. In addition, the correlation coefficients (r = -0.1007, r = -0.1189) for effect size (corrected, uncorrected) and sample size also represent the lack of relationship between effect size and sample size. Effect sizes offer a clearer picture of relationships present in the data; these indices are not clouded by sample size like statistics generated from statistical significance testing. Indeed, it is discouraging that only 5 researchers in the articles presented this information. Fagley and McKinney (1983) stated, "We feel that an understanding and

use of indices of effect size would prevent statistical tests from being misinterpreted as indicators of importance" (p. 299). Certainly, psychological research such as that reviewed in this paper could be better interpreted by the use of effect sizes.

In summary, the present paper discussed common criticisms of statistical significance testing from both historical and contemporary perspectives. Variance-accounted-for effect sizes were presented as an alternative to statistical significance testing. A review of published research in a psychological journal revealed a continued reliance on statistical significance testing on the part of researchers. Finally, scatterplots and correlation coefficients were used to illustrate the lack of linear relationship between sample size and effect size.

References

American Psychological Association (1994). Publication manual of the American Psychological Association (4th ed.). Washington, DC: Author.

Atkinson, D. R., Furlong, M. J., & Wampold, B. E. (1982). Statistical significance, reviewer evaluations, and the scientific process: Is there a (statistically) significant relationship?. Journal of Counseling Psychology, 29, 189-194.

Bakan D. (1966). The test of significance in psychological research. Psychological Bulletin, 66, 423-437.

Carver, R. P. (1978). The case against statistical significance testing. Harvard Educational Review, 48, 378-399.

Carver, R. P. (1993). The case against statistical significance testing, revisited. Journal of Experimental Education, 61, 287-292.

Cohen, J. (1990). Things I have learned (so far). American Psychologist, 45, 1304-1312.

Cohen, J. (1994). The earth is round ($p < .05$). American Psychologist, 49, 997-1003.

Craig, J. R., Eison, C. L., & Metze, L. P. (1976). Significance tests and their interpretation: An example utilizing published research and omega squared. Bulletin of the Psychonomic Society, 7, 280-282.

Fagley, N. S., & McKinney, I. J. (1983). Reviewer bias for statistically significant results: A reexamination. Journal of Counseling Psychology, 30, 298-300.

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. Psychological Bulletin, 82, 1-20.

Hays, W. L. (1973). Statistics for the social sciences (2nd ed.). New York: Holt, Rinehart, & Winston.

Huberty, C. J. (1987). On statistical testing. Educational Researcher, 16(8), 4-9.

Huberty, C. J. (1993). Historical origins of statistical testing practices: The treatment of Fisher versus Neyman-Pearson views in textbooks. Journal of Experimental Education, 61, 317-333.

Knapp, T. R. (1978). Canonical correlation analysis: A general parametric significance testing system. Psychological Bulletin, 85, 410-416.

Kupfersmid, J. (1988). Improving what is published. American Psychologist, 43, 635-642.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. Journal of Consulting and Clinical Psychology, 46, 806-834.

Rosenthal, R. (1991). Effect sizes: Pearson's correlation, its display via the BESD, and alternative indices. American Psychologist, 46, 1086-1087.

Rosnow, R. L. & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. American Psychologist, 44, 1276-1284.

Snyder, P., & Lawson, S. (1993). Evaluating results using corrected and uncorrected effect size estimates. Journal of Experimental Education, 61, 334-349.

Thompson, B. (1988). A note on statistical significance testing. Measurement and Evaluation in Counseling and Development, 20, 146-148.

Thompson, B. (1989a). Asking "what if" questions about significance tests. Measurement and Evaluation in Counseling and Development, 22, 66-68.

Thompson, B. (1989b). Statistical significance, result importance, and result generalizability: Three noteworthy but somewhat different issues. Measurement and Evaluation in Counseling and Development, 22, 2-5.

Thompson, B. (Ed.). (1993). Statistical significance testing in contemporary practice. The Journal of Experimental Education, 61(4).

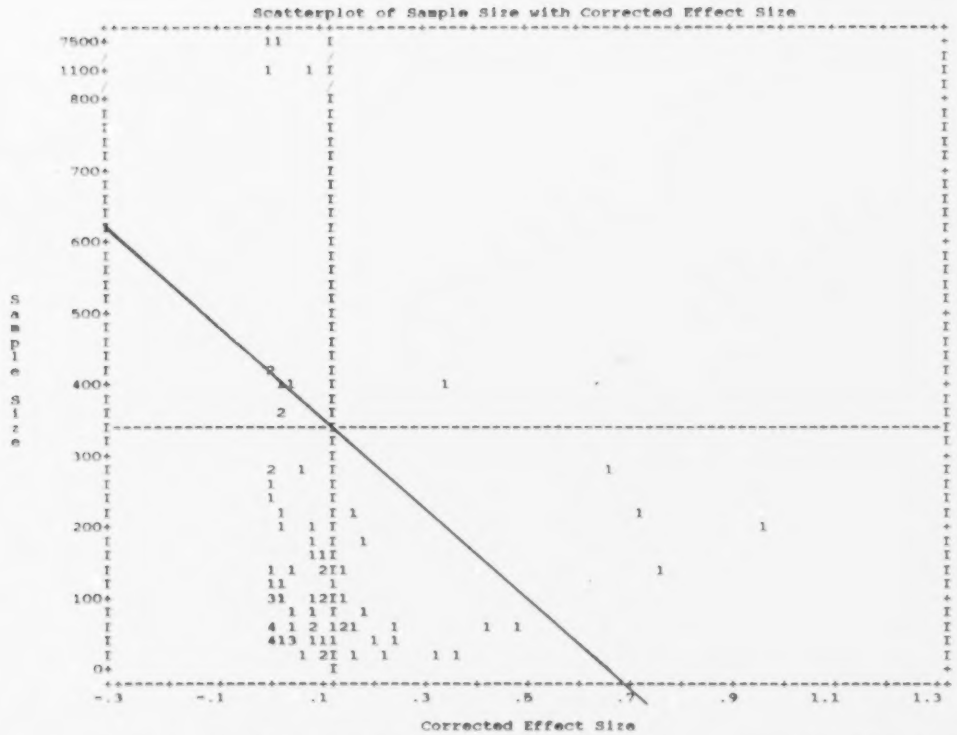
Thompson, B. (1994). The concept of statistical significance testing (An ERIC/AE Clearinghouse Digest #EDO-TM-94-1). Measurement Update, 4(1), 5-6. (ERIC Document Reproduction Service No. ED 366 654)

Wherry, R. J. (1931). A new formula for predicting the shrinkage of the coefficient of multiple correlation. Annals of Mathematical Statistics, 2, 440-451.

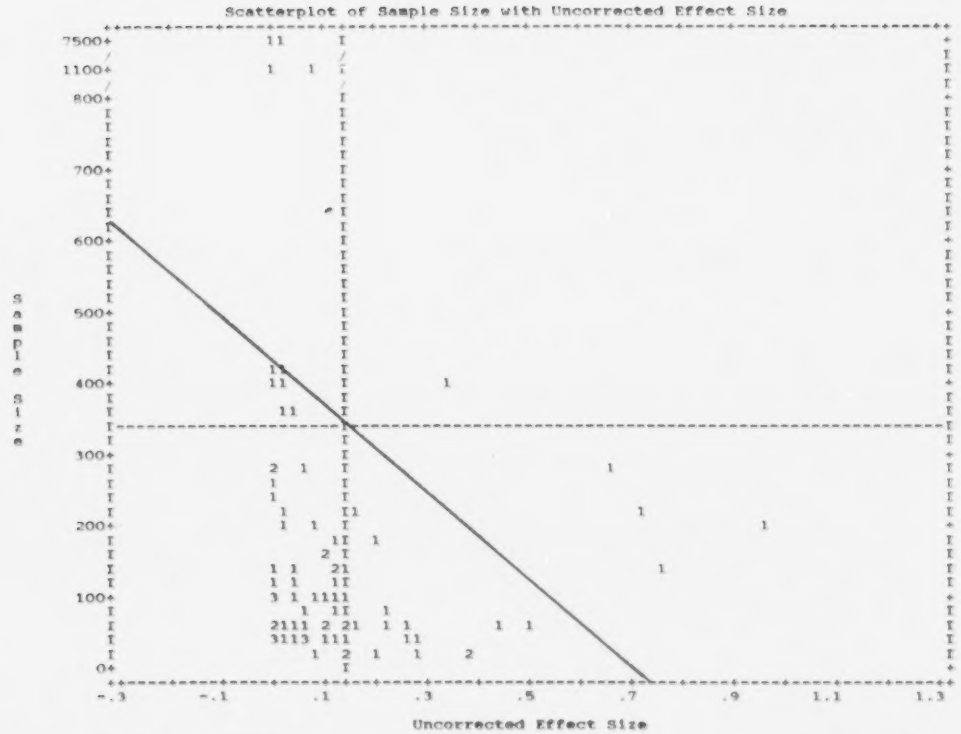
Figure Captions

Figure 1. Scatterplot of sample size and corrected effect size (omega^2) using 81 cases.

Figure 2. Scatterplot of sample size and uncorrected effect size (eta^2) using 81 cases.



Note. YHAT = 417.53 - (647.51 * corrected effect size); r = -0.1007.



Note. YHAT = 442.73 - (758.39 * uncorrected effect size); r = -0.1189.